基于神经网络与领域知识的外交国际合作元素抽取 *

张子靖 [a,b]，万常选 [a,b]，刘德喜 [a,b]，刘 玉 [a,b]，刘喜平 [a,b]，江腾蛟 [a,b]

(江西财经大学 a.信息管理学院; b.数据与知识工程江西省高校重点实验室，南昌 330013)

摘 要：为了能够实时了解国际双边合作中有价值的信息，高效地智能提取 Web 外交新闻中的国际合作元素就显得至关重要。将国际合作元素抽取抽象为类似命名实体识别的问题。首先，界定国际合作元素的内涵；其次，提取了蕴涵领域知识的规则；再次，结合神经网络与领域知识提出了面向外交新闻文本的国际合作元素抽取方法；最后，在相同语料库中与神经网络方法以及自身规则组合进行了比较，实验结果表明该方法具有更好的效果。

关键词：国际合作元素；神经网络；序列标注；命名实体识别；Web 外交新闻

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.08.0626

Extraction of diplomatic international cooperation elements based on neural networks and domain knowledge

Zhang Zijing[a,b], Wan Changxuan[a,b], Liu Dexi[a,b], Liu Yu[a,b], Liu Xiping[a,b], Jiang Tengjiao[a,b]

(a. *School of Information Technology*, *b.Key Laboratory of Data & Knowledge Engineering*, *Jiangxi University of Finance & Economics*, *Nanchang* 330013, *China*)

Abstract: In order to get valuable information in bilateral cooperation in real time, it is of utmost importance to efficiently extract international cooperation elements in Web diplomacy news. This paper **abstracte**d international cooperation element extraction into a problem similar to named entity recognition. First of all, it defined the connotations of international cooperation elements.Secondly, it extracted the rules that contained domain knowledge.Then it proposed a framework for extracting international cooperation elements for diplomatic news texts which combined with neural networks and domain knowledge.Finally, the method was compared with the neural network method and its own rule combination in the same corpus. The experimental results show that the proposed method has better results.

Key words: international cooperation elements; neural networks; sequence labeling; named entity recognition; Web diplomatic news

0 引言

国际合作元素抽取是自然语言处理(natural language processing, NLP)研究的一个分支领域。抽取方法需要能够智能识别外交新闻文本中出现的国际合作元素，如“一带一路”、农产品深加工、达特卡—克明输变电工程、《中国加入世贸组织议定书》、斯图加特德中友好协会等。研究者可以在此基础上进一步深入挖掘中国国际合作的产业结构、共性产业、新兴产业、优势产业、特色产业、产业合作倾向、产业合作成效、产业迁移和产业合作短板等，从而实现中国国际合作情况的知识发现，为走出去企业提供实时的中国国际合作信息服务，避免中国企业走出去的盲目性。随着网络新闻的发展，以 Web 为载体的外交新闻(简称为 Web 外交新闻)具有真实性、权威性、广泛性、时效性特点。透过中国 Web 外交新闻这一窗口抽取国际合作元素，可以为中国国际合作情况的知识发现提供新的研究渠道和必要的技术支持。

本文把国际合作元素抽取问题抽象为类似于命名实体识别(named entity recognition, NER)任务。传统的 NER 任务的目标是识别出非结构化文本中的信息单元，包括人名、地名和机构名以及数值表达式(如时间、日期、金额和百分比)等。NER 被看做是语言学中的序列标注任务，类似的任务有分词、词性标注以及机器翻译等。

大部分传统且表现较好的序列标注模型是线性统计模型，包括隐马尔可夫场(hidden Markov model, HMM)和条件随机场(conditional random field, CRF)[1,2]。它们的效果受到人工构建的特征和任务数据集本身特点的影响，如 NER 的效果会受到分词结果中词性特征的影响。然而，由于人工特征构建过程所需工作量和专业知识较大，使得对此类方法的改进出现了瓶颈。近年来，为了克服传统模型的局限性，随着词向量的出现，非线性的神经网络(neural network, NN)模型被广泛运用到 NLP 问题中，在 NER 任务的处理中已取得了与传统模型方法相媲美的结果[3~5]。

通过对中国 Web 外交新闻的阅读和分析，本文提出了国际合作元素的概念，分析了国际合作元素的特点，提取了涵含领域知识的规则；并在此基础上提出了抽取国际合作元素的基本策略：首先，从 Web 中获取中国外交新闻文本，进行分词处理，并对分词序列进行人工标注；然后，基于 Web 外交新闻和中文维基百科组成的数据集，训练一个词向量模型；

收稿日期：2018-08-22；**修回日期：**2018-10-01　**基金项目：**国家自然科学基金资助项目（61562032，61662027，61762042，61462037）；江西省自然科学基金资助重大项目（20152ACB20003）

作者简介：张子靖（1994-），男，江西省南昌市人，硕士研究生，主要研究方向为信息抽取、数据挖掘；万常选（1962-），男（通信作者），教授，博导，博士，主要研究方向为 Web 数据管理、情感分析、信息检索、数据挖掘（wanchangxuan@263.net）；刘德喜（1975-），男，教授，博导，博士，主要研究方向为自然语言处理、信息检索、Web 数据管理；刘玉（1993-），女，硕士研究生，主要研究方向为信息抽取、数据挖掘；刘喜平（1981-），男，副教授，主要研究方向为信息检索、数据挖掘；江腾蛟（1976-），女，讲师，博士，主要研究方向为情感分析、数据挖掘、Web 数据管理.

再次，基于词向量模型通过 BiLSTM-CNNs-CRF 神经网络结构训练国际合作元素抽取模型，并完成初步抽取；最后，通过分析初步抽取结果，借助外部词典和已提取的规则等领域知识改进抽取结果。

1 相关研究

本节简要介绍 NLP 研究中与神经网络和序列标注任务相关的一些研究进展情况。

在 NLP 的研究中，卷积神经网络(convolutional neural networks, CNN)、递归神经网络(recurrent neural networks, RNN)和长短期记忆(long short-term memory, LSTM)网络的应用较为广泛。文献[5,6]的研究表明，CNN 模型在英文文本中可以有效地从单词的字母中提取形态学信息(如单词的前缀或后缀)，并且把它编码为神经网络的表现形式。文献[7]已经证明，RNNs 模型虽然在理论上能够获取长距离依存关系，但是实际上由于梯度消失和梯度爆炸的问题，在解决长距离依存问题上遇到了困难。

为了应付梯度消失的问题，LSTM 作为 RNNs 的变体被提出，每个 LSTM 单元由三种门来保护和控制单元状态，其中，输入门控制输入的幅度，遗忘门控制之前记忆状态的输入幅度，输出门控制最终记忆的输出幅度。LSTM 利用了过去的动态时序信息和当前时刻的输入信息预测当前时刻的输出标记。

为了充分利用时间序列信息，文献[8]提出了 Bi-LSTM (Bi-directional LSTM)。Bi-LSTM 在处理很多同时需要利用到过去上下文信息和未来上下文信息的任务时，取得了不错的效果。Bi-LSTM 的基本思想是将每个序列向前和向后呈现为两个单独的隐藏状态，以分别捕获过去和未来的信息，然后将两个隐藏状态连接起来形成最终输出。

序列标注问题是 NLP 中的研究热点问题。它的应用范围非常广泛，包括 NER、词性(part-of-speech, POS)标注、浅层句法分析和机器翻译等。

文献[9]中提出了一种区别于传统 HMM 的 iHMM(infinite HMM)方法，并把它应用于词性标注领域；应用 iHMM 解决了在无监督 Markov 模型中选择隐藏状态数目的问题；在实验过程中实现了并行 iHMM 算法，在华尔街日报数据集上应用了两个非参数化(dirichlet process 和 pitman-yor process)先验；在结果评估中采用聚类评估指标，获得了比之前的工作相同或更好的评估结果；并基于该结果，用评估无监督词性标注标记器的输出替代全监督的词性标注标记器的输出，应用于浅层句法分析任务，取得了比较好的实验结果。

文献[10]将 CRF 应用于浅层句法分析任务中，并取得了不错的效果；文中提到，基于现代优化算法的改进训练方法，对于实现这样的结果至关重要。在实验过程中对模型和训练方法进行了大范围的比较。

文献[11]提出了一种新的结合 Bi-LSTM 和 CRF 的神经网络架构，在没有利用特定语言知识和资源的前提下，在 4 种语言的 NER 任务中取得了良好的效果。在文献[11]的神经网络架构基础上，文献[12]添加了一个 CNNs 网络层，以有效地从单词的字母中提取形态学信息。

现有的机器翻译系统，无论是基于短语的还是基于神经网络的，都几乎完全依赖于具有明确分割的词级建模。文献[13]使用来自 WMT15 的平行语料库对 4 种语言对(En-Cs, En-De, En-Ru 和 En-Fi)上的词根级(subword-level)编码器和字符级(character-level)编码器进行评估；实验结果说明，字符级编码器比词根级编码器的效果更好或相当。

2 国际合作元素的内涵界定

2.1 国际合作元素的定义

定义 1 国际合作元素是指在外交新闻文本中能够直接或间接表明双方或多方合作领域以及合作意向的最大连续字符序列。国际合作元素包含规划名称(简称为规划类)、产业领域名称(简称为产业类)、项目名称(简称为项目类)、协定/协议/合作备忘录等文件名称(简称为协定类)、合作平台名称(简称为平台类)五大类别。

各类别举例如下：

例 1 近年来，双方加快发展战略对接，不断推进“一带一路”框架下各领域务实合作，产能合作也取得新进展。

规划名称：“一带一路”。

例 2 中方愿意进口更多吉尔吉斯斯坦的水果、肉类、奶粉等优势农产品，丰富中国居民的餐桌。

产业领域名称：水果、肉类、奶粉、农产品。

例 3 达特卡—克明输变电工程结束了吉尔吉斯斯坦电力资源分布不均、输送不畅的历史，南北公路竣工后将实现吉尔吉斯斯坦人民天山变通途的梦想，奥什市医院建成后将为南部地区居民提供更加优质的医疗服务。

项目名称：达特卡—克明输变电工程、南北公路、奥什市医院。

例 4 德方高度赞赏中方支持欧洲一体化，愿进一步推动欧盟履行《中国加入世贸组织议定书》有关承诺，希望欧中投资协定谈判尽快完成。

协定/协议/合作备忘录等文件名称：《中国加入世贸组织议定书》。

例 5 我宣布，中方将设立中拉产能合作专项基金，为此提供 300 亿美元融资，支持中拉产能合作项目。

合作平台名称：中拉产能合作专项基金。

在外交新闻中并非所有句子都包含国际合作元素，将包含国际合作元素的句子称为外交合作句，反之则称为非外交合作句。根据以上定义，把从外交合作句中抽取国际合作元素的问题抽象为一个类似于序列标注任务，即在已经进行过分词处理的句子中找出最符合要求的标注序列。

2.2 国际合作元素的特点

国际合作元素的抽取虽然与 NER 有相似之处，但国际合作元素本身具有很多一般命名实体不具备的特点。具体分析如下：

a)国际合作元素中可能包含标点或特殊符号，而这些标点和特殊符号经常会成为划分国际合作元素的边界标志。

例 6 为延续业已开展 35 年的合作，双方商定，在中国计量科学研究院与德国联邦物理技术研究院 2014 年 4 月签署的中德计量合作协议框架下，加强在质量、时间、温度法制计量标准领域的交流。

例 7 在此基础上，双方致力于深化工业、城镇化及农业等领域的创新合作，并在此框架下共同应对可持续发展和全球公共产品保护的任务及挑战。

例 6 中，“质量、时间、温度法制计量标准”联合起来被看做是国际合作元素中产业类别的 1 个实例，在这一实例中包括 2 个顿号。而例 7 中的工业、城镇化、农业被看做是国际合作元素中产业类别的 3 个实例，而这 3 个国际合作元素被顿号分割。

b)相同的国际合作元素在不同的语境下可能被划分为不同的类别。

例 8　在今年 6 月召开的中阿合作论坛第六届部长级会议上，　　主席提出了中国同阿拉伯国家共建“丝绸之路经济带”和“21 世纪海上丝绸之路”的宏伟构想。

例 9　　表示，中方建设 21 世纪海上丝绸之路的倡议同印尼发展战略有契合之处。

例 8 中，21 世纪海上丝绸之路被看做是国际合作元素中平台类别的 1 个实例。而例 9 中，将 21 世纪海上丝绸之路分类为国际合作元素中规划类别的 1 个实例，则更符合语境。

c) 内涵更大的国际合作元素中可能包含内涵更小的国际合作元素。

例 10　不少非洲国家领导人都提出扩大中非高速公路合作、建设高速公路网的愿望，中方对此给予积极支持，愿与非方加强合作，促进非洲高速公路逐步连接成网。

例 10 中，高速公路、高速公路网能够被看做是国际合作元素中产业类别的两个实例，其中，高速公路的内涵更小，它包含于内涵更大的高速公路网中。

由于国际合作元素的以上特点，抽取结果中会出现以下部分情况：

a)本应对称出现的标点符号在抽取结果中未成对出现，例如：()、《》以及双引号，本文将此类现象称为“不规范抽取”问题；

b)抽取结果仅仅是真实结果中的子部分，同时抽取结果与真实结果在国际合作元素的分类上相同，本文将此现象称为“未完全抽取”问题；

c)在初步抽取过程中，真实结果中的 1 个国际合作元素的全部或部分可能被分割为 2 个或多个国际合作元素，同时分割产生的国际合作元素与真实结果的类别相同，本文将此现象称为“分割抽取”问题；

d)抽取的国际合作元素完全正确，但在分类时出现错误，即“分类错误”问题；

e)真实结果中的国际合作元素并未在抽取结果中出现，即“完全未抽取”问题；

f)与 e)相反，抽取结果中出现的国际合作元素并不是真实的国际合作元素，即“完全抽取错误”问题。

3　领域知识的提取

针对 2.2 节中提到的国际合作元素本身的特点，本文发现并提取了以下涵含领域知识的规则：

规则 1　抽取结果中，()、《》以及双引号必须成对出现，如果抽取结果收到影响，必须对抽取结果作出相应调整。

规则 2　开始词和结束词为出现在国际合作元素边缘但在初步抽取中未被抽取的词语。如果满足规则 1 的抽取结果 r 至少包含 2 个词，则在左边补充开始词(B)、右边补充结束词(E)。如果 BrE，Br，rE 之一出现在原句子中，则把扩充后的结果作为新的抽取结果；如果出现不止 1 类，则取长度较长者作为结果。

规则 3　中间词为出现在较长的国际合作元素中但在初步抽取中“分割”了国际合作元素的词语。记 $R_i=\{r_{i,j}\,|\,1\le i\le m, 1\le j\le n_i, r_{i,j}$ 是满足规则 1 并已经过规则 2 处理的第 i 个句子中包含的第 j 个国际合作元素，m 为句子个数，n_i 为第 i 个句子中包含的国际合作元素个数$\}$，对于 R_i 中每个至少包含 2 个词的国际合作元素，按照出现位置把与其相邻的国际合作元素从左到右和其两两组合，并在其中插入中间词(M)，即 $r_{i,j}\text{M}r_{i,j+1}$，不考虑中间词(M)的情况下则为 $r_{i,j}r_{i,j+1}$，统一记为 $r_{i,j}(\text{M})r_{i,j+1}$。如果 $r_{i,j}(\text{M})r_{i,j+1}$ 出现在原句中，则把 $r_{i,j}(\text{M})r_{i,j+1}$ 替代 $r_{i,j}r_{i,j+1}$ 作为新的抽取结果。

为了更好地表述规则 4，先定义触发词和边界标志如下。

定义 2　触发词是与某类国际合作元素搭配出现或包含在国际合作元素内，并可用于判别国际合作元素类别的词语或词组。根据触发词出现的位置，可把触发词分为 3 类：出现在国际合作元素左侧的称为前触发词，出现在国际合作元素右侧的称为后触发词，出现在国际合作元素内部的称为内触发词。

定义 3　边界标志是用以决定前触发词、后触发词与国际合作元素共现窗口大小的符号或词语，多为标点符号，少数情况为长句中能缩小范围的词语。共现窗口是指边界标志首端到国际合作元素末端(前窗口)或国际合作元素首端到边界标志末端(后窗口)。

规则 4　对于每个抽取结果，在其前窗口查找前触发词，若找到，则改变抽取结果的类别；反之，则保留原结果。后窗口同理。内触发词则在国际合作元素中查找。三种触发词的优先级(即决定权)由高到低依次为内触发词、前触发词、后触发词。

对于以上四条规则，规则 1 是为了解决初步抽取结果中出现的“不规范抽取”问题，规则 2、规则 3 和规则 4 则是利用领域知识、通过不同的处理方式分别解决初步抽取结果中出现的“未完全抽取”“分割抽取”“分类错误”3 类问题。

最后，对于“完全未抽取”“完全抽取错误”2 类问题，由于缺乏必要的线索，无法在初步抽取结果的基础上通过领域知识而解决，因此不属于本文研究的范畴。

4　国际合作元素的抽取方式

国际合作元素的抽取与序列标注任务具有相似的特点；但是相比传统序列标注任务，国际合作元素抽取任务涉及的类别较多，且长度分布不均匀，这使得一般的序列标注方法抽取得到的结果仍然有改进的余地。从元素类别来说，国际合作元素的类别较多且类别的分类以“语义”为标准，相比传统的人名地名识别的分类更为困难；从元素长度而言，国际合作元素的长度(词语个数)分布不均匀，如例 6 中的“质量、时间、温度法制计量标准”不仅包含多个词语而且还包含标点符号。为了兼顾以上两点，本文的抽取策略是：首先，采用近期在序列标注任务上表现优异的神经网络模型对国际合作元素进行初步抽取；然后，将初步抽取的结果作为输入，借助已提取的领域知识规则对抽取结果进行优化，并把优化结果作为最后的输出结果。方法框架如图 1 所示，对应算法 1 和 2。

算法 1　基于神经网络的国际合作元素抽取

输入：已训练好的神经网络模型，已进行分句和分词处理的文件。

输出：初步抽取结果。

```
for 测试集中的一个批次(batch)的每一个句子(s) {
    for 句子(s)中的每一个词语(x_t) {
        /* 在词向量文件(WV)中查找词语 x_t 所对应的词向量表示 */
        word_embedding ← WV[x_t];
        /* 使用 w2c 方法计算词语 x_t 的字符表示 */
        char_representation ← w2c(x_t);
        /* 拼接 word_embedding 和 char_representation 作为词语 x_t 的神
           经网络的输入 */
        embedding ← concat(word_embedding, char_representation);
        /* 前向过程：决定什么信息应该被神经元遗忘 */
        ff_t ← σ(W_ff ·[h_{t-1}, x_t]+b_ff);
        /* Sigmod 层决定要更新的数值 */
```

$fi_t \leftarrow \sigma(W_{fi} \cdot [h_{t-1}, x_t] + b_{fi})$;

/* tanh 层生成新的候选数值， fC_t 会被增加到神经元状态中 */

$fC_t \leftarrow \tanh(W_{fC} \cdot [h_{t-1}, x_t] + b_{fC})$;

/* 更新旧的神经元状态 fC_{t-1} 到新的神经元状态 fC_t */

$fC_t \leftarrow ff_t * fC_{t-1} + fi_t * fC_t$;

/* 使用 sigmod 层决定哪一部分的神经元状态需要被输出 */

$fo_t \leftarrow \sigma(W_{fo} \cdot [h_{t-1}, x_t] + b_{fo})$;

/* 让神经元状态 fC_t 经过 tanh（使输出值变为-1~1 之间）层，并乘以 sigmod 门限的输出 fo_t 后作为前向过程的最终结果 */

$fh_t \leftarrow \tanh(fC_t)\, fo_t$;

/* 后向过程：计算方法与前向过程相同，只是计算方向相反；此处只给出后向过程的最终结果 */

$bh_t \leftarrow \tanh(bC_t)\, bo_t$;

/* 拼接前向过程与后向过程的结果，再用模型中的权重矩阵 ***W*** 和偏置矩阵 ***b*** 计算出未解码的 ***output*** */

output ← concat(fh_t , bh_t)***W*** + ***b***;

}

/* 使用条件随机场进行解码，***trans_params*** 为模型中的转移矩阵，***pre_tags*** 为句子对应的标签序列 */

pre_tags ← crf_decode(***output***, ***trans_params***);

}

图 1　抽取国际合作元素的方法框架

Fig. 1　Methodological framework for extracting international cooperation elements

算法 2　基于领域知识的优化

输入：一个句子的神经网络的抽取结果(词语-标记)，开始结束词集合，中间词集合，边界标志集合以及触发词集合。

输出：最终抽取结果(词语-标记)。

```
/* 1)规范化 */
while (句子 sentence 中找到了未匹配的“《》”，“()”或双引号) {
    /* 查找句子 sentence 中未匹配的“《》”，“()” 或双引号，并返回第一次出现的位置 */
    spot ← symbolmatch(sentence);
    if (未匹配的符号被标记为“其他”) {
        删除该词语-标记对;
    }
    else if (未匹配的符号被标记为国际合作元素的“开始”) {
        找到对应元素的末尾，添加对应符号，并修改标记;
    }
    else {
        找到对应元素的开始，添加对应符号，并修改标记;
    }
}
/* 2)扩充元素 */
for 句子中的每一个国际合作元素 {
    if (国际合作元素至少包含两个分词) {
        if (国际合作元素的前一个词出现在开始词集合中，且后一个词出现在结束词集合中) {
            NE ← B + r + E;
        }
        else if (国际合作元素的前一个词出现在开始词集合中) {
            NE ← B + r;
        }
        else if (国际合作元素的后一个词出现在结束词集合中) {
            NE ← r + E;
        }
    }
}
/* 3)合并元素 */
for 句子中的每一个国际合作元素 {
    if (国际合作元素至少包含两个分词) {
        if (该国际合作元素之后仍存在其他国际合作元素) {
            M ← 该国际合作元素与下一个国际合作元素中间的部分;
            if (M 出现在了中间词集合中) {
                NE ← r_i + M + r_{i+1};
            }
        }
    }
}
/* 4) 修正分类 */
找出句子中包含的所有边界标志;
通过边界标志和国际合作元素的位置明确句子中的窗口位置和大小;
在每个窗口中查找触发词;
根据触发词的优先级更新国际合作元素的类别;
```

4.1　神经网络层的训练

在神经网络层中，采用 CNNs-BiLSTM-CRF[12]结构获取初步的序列标注结果。其中，CNNs 层是利用每个词中的每个字对应的字向量，通过 CNNs 结合后得到一个词的字符表示(char representation)，其网络结构如图 2 所示；BiLSTM 层的输入是每个词的词向量表示(word embedding)以及其对应的字符表示的连接，输出为每一个词对应的标注状态；对于

序列标注任务，CRF 层在考虑邻域中标签之间的相关性、并对给定输入句子的最佳标注链进行联合解码时的帮助很大。

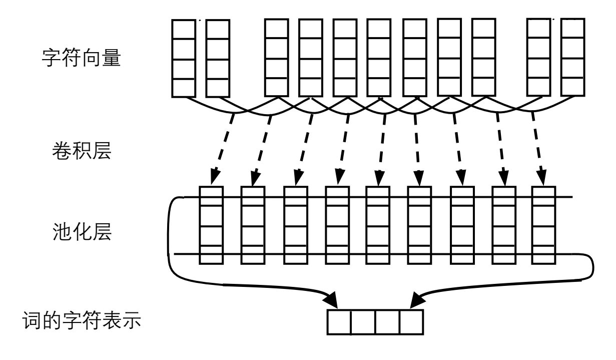


图 2　得到词的字符表示的 CNN 网络结构

Fig. 2　CNN network structure for character representation of words

在 CNNs-BiLSTM-CRF 模型的训练过程中，本文所使用的参数与文献[12]中略有不同。具体的参数选择情况如表 1 所示。其中，adam 为学习率(learning rate)优化时所采用的一种利用一阶导数的优化算法，此类算法的选择会影响到 batch size 在数量级上的选择，在综合各类因素后 batch size 最终取值为 20；gradient clipping 取值为-1(小于 0)表示不使用 gradient clipping。

表 1　神经网络层训练的参数选择

Table 1　Parameter selection of neural network layer training

层	超参数	值
CNN	window size	3
	number of filters	30
LSTM	state size	300
	initial state	0
	peepholes	No
dropout	dropout rate	0.5
Other parameter	batch size	20
	initial learning rate	0.001
	decay rate	0.9
	gradient clipping	-1
	learning rate method	Adam

4.2　基于领域知识的优化

在分析了由神经网络层抽取到的初步结果后，本文发现初步抽取结果中仍然存在着 2.2 节中所提到的 6 类问题。其中，造成“不规范抽取”问题的原因有两个，一是数据集本身出现特殊符号未配对现象；二是抽取过程中丢失符号。此类问题是底层的抽取错误，可能会对后续优化过程产生影响，因此必须首先对初步抽取结果进行规范化处理。在规范化处理中，要保证抽取结果中特殊符号的出现必须与数据集中包含的特殊符号“同步”，即如果数据集中特殊符号配对出现，则抽取结果中特殊符号也必须配对出现；否则，抽取结果中不出现特殊符号。规范化处理对应于第 3 节中的规则 1。

对于“未抽取完全”和“分割抽取”2 类问题，虽然它们都已经抽取到了真实结果的组成部分，但是这样的结果并不完整，并不是真正的国际合作元素。然而，正因为它们抽取到了真实结果的组成部分，才提供了针对初步抽取结果进行“扩充”与“合并”来优化抽取结果的机会。“扩充”与“合并”的具体策略对应于第 3 节中规则 2 和规则 3。

在优化了“未抽取完全”和“分割抽取”两类问题之后会产生新的国际合作元素，而在新国际合作元素和原国际合作元素中仍然会存在“分类错误”问题。为此本文提出了相应的优化策略，对应于第 3 节中的规则 4。

5　实验结果与分析

5.1　实验数据集和环境

实验部分使用到的第 1 部分数据是中国 Web 外交新闻数据集，它来源于中国外交部网站[1]提供的中国驻各国使馆新闻文本。首先，针对获取的中国 Web 外交新闻数据集，利用哈工大提供的 LTP 分词工具[2]依次进行分句、分词处理；然后，根据任务需求对分词结果进行人工标注。本文选用 3 个人作为数据集的标注者，以少数服从多数决定正确答案；当出现无法决定的情形，则由 3 人讨论并经本文确认标注结果。3 个人的组成为 2 位本科生和 1 位研究生，且都有财经基础课程学习经历。

依据句子中是否包含国际合作元素，可分为外交合作句和非外交合作句。中国 Web 外交新闻数据集中共有 10 030 个外交合作句，可作为抽取国际合作元素的数据集，记为 $dataSet_1$。实验部分使用到的第 2 部分数据是中文维基百科的网页数据集，记为 $dataSet_2$。

在实验过程中，训练集、验证集和测试集的划分如下：将 $dataSet_1$ 中的 8 000 个、1 000 个和 1 030 个句子分别作为训练集、验证集和测试集。在训练集、验证集和测试集中，人工标注的 5 种类别国际合作元素的统计数据如表 2 所示。

表 2　数据集中不同类别国际合作元素的统计结果

Table 2　Statistical results of different types of international cooperation elements in the dataset

类别	训练集		验证集		测试集		合计	
	数量/个	占比/%	数量/个	占比/%	数量/个	占比/%	数量/个	占比/%
规划	516	3.36	62	3.11	132	3.28	710	3.32
产业	10087	65.63	1333	66.78	2690	66.89	14110	65.97
项目	1655	10.77	228	11.42	426	10.59	2309	10.80
协定	618	4.02	66	3.31	151	3.75	835	3.90
平台	2493	16.22	307	15.38	623	15.49	3423	16.01
合计	15369	100	1996	100	4022	100	21387	100

在第 3 节中提到的开始词、结束词、中间词、触发词以及边界标志均从人工构建的词典中获得。其中，开始词和结束词分别提取了 46 个和 62 个词语；中间词共提取了 12 个词语；触发词共提取了 27 个词语；边界标志共提取了 4 个标点符号和 6 个词语。

实验环境：HP Z840 图形工作站，采用 Python 语言，并使用谷歌提供的 TensorFlow 深度学习框架。实现方案：参考了 TensorFlow 训练神经网络模型的标准机制，即网络结构设计与数据流独立；在神经网络初步抽取之后，根据提取的领域知识，逐个优化领域知识的实现。

5.2　实验设计与实验结果分析

本节首先对第 3 节中提取出的领域知识在数据集中的影响范围进行分析；然后再针对本文方法与 CNNs-BiLSTM-CRF[12]模型抽取国际合作元素的精确率 *P*、召回率 *R* 以及 *F*1 值进行实验对比分析。

1)领域知识的提取效果

领域知识的提取效果如表 3 所示，分析结果是基于 $dataSet_1$ 进行的。其中，规则 1 用于规范化初步抽取结果，规则 2 和规则 3 用于弥补由“未抽取完全”和“分割抽取”问题带来的抽取错误，规则 4 则用于改进“分类错误”问题。

表 3　领域知识的提取效果分析

[1]http://www.fmprc.gov.cn/web/zwjg_674741/zwsg_674743/yz_674745

[2] http://www.ltp-cloud.com/demo/

Table 3 Analysis on extraction effect of domain knowledge

规则	问题数量/个	发现数量/个	纠正数量/个	发现率/%	纠正率/%
规则 1	24	24	23	100	95.83
规则 2	448	501	389	89.42	77.64
规则 3	92	118	81	77.97	68.64
规则 4	601	738	549	81.44	74.39

由于 4 条规则均是为优化神经网络层的抽取结果而提出的，前题是保证在不影响神经网络层已抽取正确结果的前提下，尽可能准确地弥补或改进错误，因此一般的规则评价标准并不适用于此类情况。于是本文提出用如下两个指标来评价领域知识的提取效果。

发现率=真实问题个数/发现问题个数*100%

纠正率=纠正问题个数/发现问题个数*100%

其中，发现率用来评测规则是否能够尽可能多地发现错误结果，而不去影响正确结果；纠正率用来评价规则所发现的错误是否能够被尽可能多地纠正。

2)国际合作元素的抽取效果

具体的实验过程如下：首先，利用 $dataSet_1$ 和 $dataSet_2$ 训练出一个 300 维的词向量；然后，将词向量、训练集和验证集作为输入，通过网络模型训练出一个国际合作元素的标注模型；最后，可以评估出测试集在标注模型上的表现，如表 4 所示。本文方法的实验结果如表 5 所示。

表 4 CNNs-BiLSTM-CRF 模型的实验结果

Table 4 Experimental Results of the CNNs-BiLSTM-CRF model

类别	*P*/%	*R*/%	*F*/%
规划	88.33	77.94	82.81
产业	88.75	90.06	89.40
项目	69.47	66.00	67.69
协定	72.83	79.76	76.14
平台	83.02	85.13	84.06
合计	85.32	86.08	85.70

表 5 本文方法的实验结果

Table 5 Experimental results of the proposed model

规则	类别	*P*/%	*R*/%	*F*/%
规则 1	规划	90.00	79.41	84.38
	产业	88.75	90.06	89.40
	项目	69.47	66.00	67.69
	协定	72.83	79.76	76.14
	平台	83.18	86.08	84.60
	合计	85.39	86.28	85.83
规则 2 规则 3	规划	90.16	80.88	85.27
	产业	89.10	90.28	89.69
	项目	72.73	72.00	72.36
	协定	78.49	86.90	82.49
	平台	85.54	87.97	86.74
	合计	86.51	87.66	87.08
规则 4	规划	88.52	79.41	83.72
	产业	88.70	90.13	89.41
	项目	69.79	67.00	68.37
	协定	74.19	82.14	85.27
	平台	84.47	86.08	84.06
	合计	85.60	86.53	86.06
所有规则结合	规划	90.00	92.65	91.30
	产业	89.27	90.65	89.95
	项目	75.36	79.50	77.37
	协定	80.21	91.67	85.56
	平台	87.35	91.77	89.51
	合计	87.16	89.83	88.48

为了综合并直观地对所有的实验结果进行对比分析，将本文方法与 CNNs-BiLSTM-CRF 模型的实验结果(考虑 *F*1 值)通过柱状图进行比较直观的比对分析，结果如图 3 所示。

从图 3 可以看出，本文采用基于领域知识的优化策略，对提高 CNNs-BiLSTM-CRF 模型的 *F*1 值有明显的效果。规划类别的 *F*1 值提高了 8.49 个百分点，提高率为 10.25%；产业类别的 *F*1 值提高了 0.55 个百分点，提高率为 0.62%；项目类别的 *F*1 值提高了 9.68 个百分点，提高率为 14.30%；协定类别的 *F*1 值提高了 9.42 个百分点，提高率为 12.37%；平台类别的 *F*1 值提高了 5.45 个百分点，提高率为 6.48%。

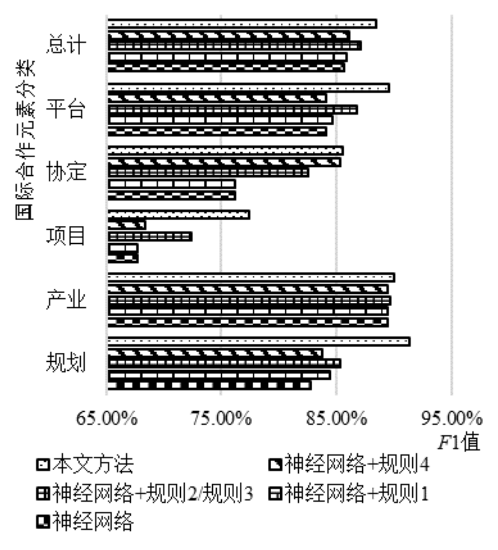


图 3 采用基于领域知识的优化策略对提高 CNNs-BiLSTM-CRF 模型 *F*1 值的对比分析

Fig. 3 Comparative analysis for improving *F*1 of the cnns-bilstm-CRF model using the optimization strategy based on domain knowledge

从产业类别来看，领域知识对该类别的影响最小，使得领域知识在该类别上的优化效果很低，原因主要有 3 点：一是产业类别的国际合作元素的词长都较短，因此 CNNs-BiLSTM-CRF 模型对产业类别的国际合作元素的抽取效果较好(*F*1 值为 89.40%)；二是在数据集中产业类别的国际合作元素总量较多(占总量的 65.97%)，CNNs-BiLSTM-CRF 模型对于产业类别的国际合作元素的训练效果较好；三是“扩充”与“合并”规则主要是用于优化较长的国际合作元素(要求词长不小于 2)的抽取效果。

对于项目类别，相比其他类别而言，尽管领域知识使得抽取结果的 *F*1 值得到了最大程度的优化(*F*1 值提高了 14.30%)，但由于 CNNs-BiLSTM-CRF 模型对项目类别抽取的 *F*1 值是最低的(*F*1 值为 67.69%)，导致优化之后对项目类别抽取的 *F*1 值仍然是最低的(*F*1 值为 77.37%)。本文认为主要原因有 2 点：一是 CNNs-BiLSTM-CRF 模型对项目类别的召回率明显偏低(*R* 值为 66.00%)，而领域知识是基于对该模型召回的国际合作元素进行优化，尽管在优化过程中还会发现新的国际合作元素，一定程度上提高召回率，但优化之后对项目类别的召回率还是偏低(*R* 值为 79.50%)；二是项目类别的国际合作元素相较于其他类别的构成更为复杂，导致该模型的抽取精度也明显偏低(*P* 值为 69.47%)，优化之后对项目类别的抽取精度还是偏低(*P* 值为 75.36%)。

对于平台类别，由于被分类未平台类别多为机构名称，而机构名称的抽取也是传统命名实体识别的任务之一，并且从表 2 中可以看出平台类别在所有元素类别中的占比排名第二(16.01%)。因此，CNNs-BiLSTM-CRF 模型对其抽取的精

度和召回率都较为出色且平均(*P* 值为 83.02%, *R* 值为 85.13%)。但是由于国际合作元素的长度分布不均匀，平台元素相比传统的机构名在抽取过程中会遇到更多的“未完全抽取”和“分割抽取”现象。

从规划和协定类别来看，由于这些类别的国际合作元素一般都较长，同时这两类元素在数据集中的占比很小(分别为3.32%和 3.90%)，所以 CNNs-BiLSTM-CRF 模型在这 4 类国际合作元素的抽取结果上表现不佳，而领域知识则有效地提升了这两类国际合作元素提取的 *F*1 值。

6 结束语

本文首先利用分词工具对语料库中的 Web 外交新闻文本进行分词，再利用神经网络对文本中的国际合作元素进行初步抽取，最后结合人工提取的领域知识对初步抽取结果进行优化得到国际合作元素的最终抽取结果。在实验阶段，把本文方法与神经网络方法在相同语料库上进行了对比分析，验证了本文方法能取得更好的效果；同时对人工提取的领域知识在语料库上的效果进行了对比和分析。

从实验分析中可以发现，本文方法对于初步抽取结果有着较强的依赖性，所以如何提高初步抽取性能是未来工作的重点之一，尤其是对项目类别国际合作元素的抽取。其次，本文使用的语料库中五类国际合作元素的分布倾斜度较高，今后会考虑对数据集进行扩充并构建一个 5 类合作元素分布倾斜度较小的数据集，在新的数据集上对方法进行实验和改进。最后，利用从 Web 外交新闻中抽取到的国际合作元素继续开展知识发现方面的研究。

参考文献:

[1] Luo Gang, Huang Xiaojiang, Lin C Y, *et al.* Joint entity recognition and disambiguation [C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 879-888.

[2] Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution [C]// Proc of the 18th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg, PA: ACL, 2014: 78-86.

[3] Hu Zhiting, Ma Xuezhe, Liu Zhengzhong, *et al.* Harnessing deep neural networks with logic rules [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics，Volume 1: Long Papers. Stroudsburg, PA: ACL, 2016: 2410-2420.

[4] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2013: 6645-6649.

[5] Zadrozny B. Santos C D, Zadrozny B. Learning character-level representations for part-of-speech tagging [C]// Proc of the 31st International Conference on Machine Learning. New York: ACM Press, 2014: 1818-1826.

[6] Chiu J, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association of Computational Linguistics, 2016, 4(1): 357-370.

[7] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C]//Proc of International Conference on Machine Learning. New York: ACM Press, 2013: 1310-1318.

[8] Dyer C, Ballesteros M, Ling Wang, *et al.* Transition-based dependency parsing with stack long short-term memory [C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing ,Volume 1: Long Papers. Stroudsburg, PA: ACL, 2015: 334-343.

[9] Gael J V, Vlachos A, Ghahramani Z. The infinite HMM for unsupervised PoS tagging. [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2009: 678-687.

[10] Sha Fei, Pereira F. Shallow parsing with conditional random fields [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg, PA: ACL, 2003: 134-141.

[11] Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition [C]//Proc of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 260-270.

[12] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics,Volume 1: Long Papers. Stroudsburg, PA: ACL, 2016: 1064-1074.

[13] Chung J, Cho K, Bengio Y. A character-level decoder without explicit segmentation for neural machine translation [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016,Volume 1: Long Papers: 1693-1703.